Sains Malaysiana 44(10)(2015): 1531–1540

Application of Functional Data Analysis for the Treatment of Missing Air Quality Data

(Aplikasi Analisis Data Fungsian untuk Merawat Data Kualiti Udara yang Lenyap)

NORSHAHIDA SHAADAN*, SAYANG MOHD DENI & ABDUL AZIZ JEMAIN

ABSTRACT

In most research including environmental research, missing recorded data often exists and has become a common problem for data quality. In this study, several imputation methods that have been designed based on the techniques for functional data analysis are introduced and the capability of the methods for estimating missing values is investigated. Single imputation methods and iterative imputation methods are conducted by means of curve estimation using regression and roughness penalty smoothing approaches. The performance of the methods is compared using a reference data set, the real PM_{10} data from an air quality monitoring station namely the Petaling Jaya station located at the western part of Peninsular Malaysia. A hundred of the missing data sets that have been generated from a reference data set with six different patterns of missing values are used to investigate the performance of the considered methods. The patterns are simulated according to three percentages (5, 10 and 15) of missing values with respect to two different sizes (3 and 7) of maximum gap lengths (consecutive missing points). By means of the mean absolute error, the index of agreement and the coefficient of determination as the performance indicators, the results have showed that the iterative imputation method using the roughness penalty approach is more flexible and superior to other methods.

Keywords: Air quality; functional data; imputation; missing value; PM_{10}

ABSTRAK

Dalam kebanyakan penyelidikan termasuklah penyelidikan alam sekitar, data lenyap sering wujud dalam rekod dan telah menjadi masalah lazim terhadap kualiti data. Dalam kajian ini, beberapa kaedah imputasi yang berasaskan teknik analisis data fungsian telah dicadangkan dan kebolehan kaedah tersebut dikaji. Kaedah imputasi tunggal dan kaedah imputasi ulangan telah dijalankan dengan pendekatan penganggaran lengkuk menggunakan teknik pelicinan regresi dan teknik denda kekasaran. Prestasi kaedah-kaedah imputasi dibandingkan menggunakan data set rujukan cerapan sebenar pencemar PM_{10} yang telah direkodkan di stesen pemantau kualiti udara Petaling Jaya yang terletak di bahagian barat Semenanjung Malaysia. Untuk mengkaji prestasi kaedah imputasi yang dicadangkan, sebanyak seratus data set dijana untuk setiap enam paten data lenyap yang berbeza menggunakan data rujukan. Paten kelenyapan data disimulasi mengikut tiga jumlah nilai peratusan kelenyapan (5, 10 dan 15) dengan dua saiz maksimum panjang turutan kelenyapan (3 dan 7) (titik lenyap berturut). Dengan kaedah min ralat mutlak, indeks persetujuan dan nilai pekali penentu sebagai penunjuk prestasi, keputusan analisis kajian mendapati bahawa kaedah imputasi ulangan yang menggunakan pendekatan denda kekasaran adalah lebih fleksibel dan lebih baik daripada kaedah yang lain.

Kata kunci: Data fungsian; imputasi; kualiti udara; nilai lenyap; PM_{10}

INTRODUCTION

The availability of a complete data set is essential in various statistical analyses. However, in the context of air quality data, the problem of missing data often occurs due to various reasons, for instance, malfunction of equipment, human error and calibration process. The results of any statistical model and analysis could deteriorate when using incomplete records as an input in the analysis. Due to this fact, the estimation of missing values becomes the first priority in the data preparation process. Various replacement methods have been used to tackle the problem and are profoundly discussed in the literature; from the simple traditional method to the very sophisticated one. The concern has been put forward in many field of studies (Baraldi & Enders 2010; Malek et al. 2008; Preda et al. 2005; Police & Lasinio 2009; Smolinski & Hlawiczka 2007; Zhang 2011), including environmental research (Junninen et al. 2004; Plaia & Bondi 2006). Two general approaches for solving this problem are case deletion and imputation, which are the popular methods. Even though case deletion is the most common and simplest method, it has the disadvantage of losing information due to data reduction.

Recorded air quality data such as PM_{10}, NOx, CO, SO_2 and ozone are continuously measured data and the pattern of the recorded data is often non-linear and dependent of time. Instead of considering the measured data as continuous discrete values, the data can also be treated as a finite curve over an interval time period. Functional data analysis (FDA) consists of statistical techniques were used to analyze the curves data. The application of FDA for the air quality data has been identified. Among the areas of the application include the study on the trend, severity and the dynamic behaviour of a particular pollutant such PM_{10} or ozone (Gao & Niemeier 2008; Park et al. 2013; Shaadan et al. 2012), the study on the extreme values and the prediction of pollutant curves (Quintela-del-Rio & Francisco-Fernandez 2011), as well as outliers or anomalies detection and assessment (Martinez et al. 2014; Shaadan et al. 2015; Torres et al. 2011). Noticeably, the application of FDA for the treatment of missing values is rarely found. However, at a particular point of view, with the FDA approach, the problem of missing values could be overcome by means of curve estimation. In the FDA methods, converting discrete observed data into curves using curve estimation is the first step needed before further analysis is conducted. Several applications of curve estimation in the treatment of missing values have been used in several areas. Among the closely related research that is continuously being explored was the paper by Chen et al. (2010). The idea of curve fitting using the B-spline and the non-parametric regression namely the kernel smoothing to clean corrupted and missing electric energy consumption using real data of the British Columbia Transmission Corporation (BCTC) was incorporated in the work. Another research was done by Cao et al. (2008). These researchers had used the combination between the kernel smoothing and the nearest neighbor approach as the imputation-based method for microarray data. Remarkably, in both studies, the imputation was conducted by only using a one-step procedure, in which the non-iterative approach was considered. In the functional data analysis literature, missing data with short gaps can be reliably constructed by means of data conversion from discrete point data into a curve. Meanwhile, in the presence of long gaps, Ruggieri et al. (2013) proposed the empirical orthogonal function (EOF) procedure based on the functional principal component analysis (FPCA).

In this paper, several imputation methods are designed based on the methods for data conversion in FDA. In addition to the previous data conversion approach, the initial-based value and the iterative-based imputation methods are incorporated in the treatment procedure. Two common approaches; the regression and roughness penalty smoothing are applied. This study aimed to investigate the capability of the imputation methods to estimate the missing value. Both the single imputation methods and the iterative imputation methods are conducted in the experiment and the performance are then compared.

Data and Methods

In this study, the experimentation for missing data analysis is conducted using a real PM_{10} data set that has been recorded daily at hourly basis at the Petaling Jaya air quality monitoring station located in the west part of Peninsular Malaysia. A matrix data set with 153 days (rows) by 24 h (column) which was recorded during the southwest (SW) monsoon season in the year 2010 is considered. The data contains small percentage of missing values (1.3%) and have been replaced using the column median value of the complete available data (Acuna & Rodriguez 2004). The use of real data as the reference is considered in order to define the output of the performance criteria for better accuracy (Plaia & Bondi 2006). All the analyses are carried out using the free software R (R Development Core Team 2008).

The first step in the imputation analysis was to generate six patterns with 100 sets of incomplete data for each pattern from the reference set to guarantee the consistency of the results. Next, the replacement of the missing values will be conducted using the considered imputation methods. Finally, in order to evaluate the imputation methods, the performance indicators were computed for each of the six missing data patterns.

Fundamental Concept of Missing Treatment

Given that a few data points are missing within the 24 h period of a day curve, the aimed was to replace the missing value y_{miss} at time t with a value on the estimated curve $x(t)$ at the same time t being missing, where $\hat{y}_{miss} = x(t)$. Examples for the possible condition of the missing values are illustrated in Figure 1.

To solve the missing values, curve construction must be the first step to be considered. This is the fundamental step in FDA (Ramsay & Silverman 2006). A daily curve is actually a function defined over an interval of (1, 24). The discrete observations y_j, $j = 1, \ldots, 24$ are converted into a function $x_i(t)$, $i = 1, \ldots, n$, which allows for the evaluation of the function at any time point t_j. The estimation of daily curves represented by function $x_i(t)$ is conducted by means of a system of basis function expansion, which is a linear combination of K independent basis functions $\varphi_k(t)$, whereas the term β_k refers to the basis coefficient as follows:

$$x_i(t) = \sum_{k=1}^{K} \beta_k \varphi_k(t). \tag{1}$$

In this study, a linear combination of K number B- spline basis is used to represent the curve for a more flexible fitting. Splines are piecewise polynomials; thus, to define a spline basis, information on a set of knots or the K number of basis and the degree of polynomials is needed. To construct a daily curve, a degree three polynomial is employed in this study. The coefficients β_k are determined through the least square method by minimizing the sum of squared residuals (SSE) as follows:

$$\text{SSE} = \sum_{j=1}^{24}\left(y_j - x(t_j)\right)^2 = \sum_{j=1}^{24}\left(y_j - \sum_{k=1}^{K}\beta_k\varphi_k(t_j)\right)^2. \tag{2}$$

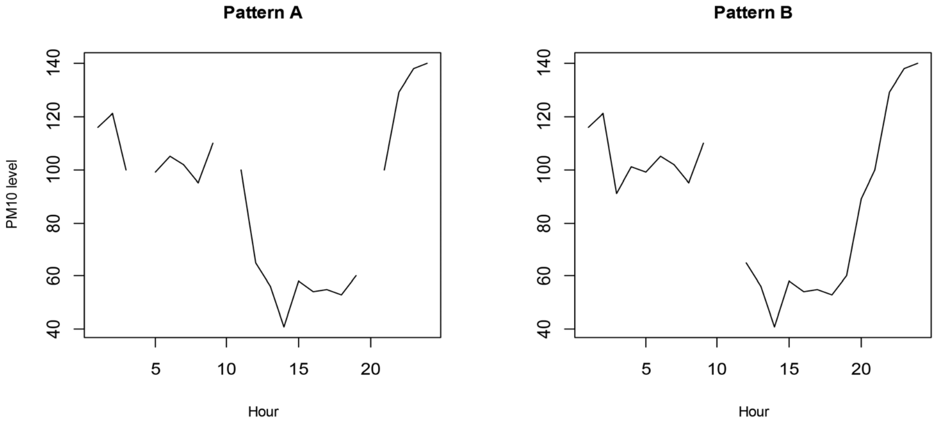

FIGURE 1. Possible patterns of missing values within a day curve

A method for smoothing a curve using the least square method is known as regression smoothing (Ramsay et al. 2009). The roughness penalty approach is an alternative approach for curve conversion. The approach allows for a finer control over the amount of smoothing. The idea of the roughness penalty approach will be incorporated into some of the methods. A measure of roughness (i.e. curvature) is defined by the square of the second derivative $(x''(t_j))^2$ and a parameter λ is used to control the roughness to prevent curve over-fitting. When the roughness is included in the fitting process, the least square fitting criterion in (2) is modified. Thus, the penalized sum of squares (PENSSE) is given as follows:

$$\text{PENSSE} = \sum_{j=1}^{24}\left(y_j - x(t_j)\right)^2 + \lambda\int\left(x''\left(t_j\right)\right)^2 dt. \qquad (3)$$

Using the roughness penalty approach, the estimate of the function is obtained by finding the function that minimizes PENSSE.

Estimation of K and λ for the missing data set For the construction of daily curves, we used an expected common K and λ as a guiding value with the assumptions that the recorded data within the data set comes from the same underpinning process. Using the available recorded data, K and λ are estimated based on the construction of the expected (mean) curve. Appropriate K is the one that gives the minimum Bayesian Information Criterion (BIC) (Huang & Shen 2004). Let m denotes the number of recorded data points (in this case m=24), K is the number of basis and SSE is the residual sum of squares or error variance of the estimated mean curve. The BIC formula is given by:

$$\text{BIC} = \log\left(\frac{\text{SSE}}{m}\right) + \frac{k}{m}\log(m). \qquad (4)$$

The common appropriate λ is determined based on the minimum value of generalized cross-validation (GCV) criterion (Craven & Wahba 1979). The criterion is defined by:

$$\text{GCV}(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{\text{SSE}}{n - df(\lambda)}\right). \qquad (5)$$

Based on the results given by the BIC and GCV analysis, we have decided to use K equals 19 and λ equals 0.00001 in this study.

GENERATION OF SIMULATED DATA SET WITH MISSING VALUES

Missing data are generated from the reference data set using six randomly simulated missing data patterns in different missing data conditions as shown in Table 1. For each generated data set, the patterns are different in complexity and are simulated with total missing data percentages of 5, 10, and 15 and according to the maximum number of consecutive missing values per rows (gap length) of three and seven. Thus, within the generated data set, different curves may have different numbers of missing values and different missing gap lengths. For example, the data set with pattern P05-G3 consists of 5% missing and the maximum gap length per day can reach up to three consecutive values.

IMPUTATION METHODS

Based on the data conversion techniques in FDA, seven imputation methods are considered. The methods are classified into three categories; the single imputation, the iterative imputation without roughness penalty, and the iterative imputation with roughness penalty approach.

Functional mean data method (FMeanDM) and functional median data method (FMedDM) These single imputation methods allow a missing value at time t_j to be replaced by the corresponding point value that lies on the mean or median curve obtained from the available data. The mean curve is the mean of the concentration level whereas the median curve is the middle value at time column t_j, j = {1,. . . , 24} across the replication of completed daily curves.

TABLE 1. Six different patterns of missing generated data

Pattern	P05-G3	P05-G7	P10-G3	P10-G7	P15-G3	P15-G7
Missing percentage	5	5	10	10	15	15
Maximum gap length	3	7	3	7	3	7

Functional mean-based iterative method (FMeanBIM) and functional median-based iterative method (FMedBIM) An iterative approach was developed for the missing data imputation to enhance the single imputation strategies. Thus, the iteration process will make full use of all the useful information, including the instances with the missing values and improving the imputation performance (Pighin & Ieronutti 2008). Initially, before the imputation started, the missing values were replaced by the point values on the mean or median at the first iteration and then the missing values were iteratively imputed until the algorithm converges. The convergence is satisfied whenever the difference in the error (MPAD) between the imputation at the current and the previous iteration is at most 0.005. MPAD is defined as the mean proportion of the absolute difference between the point of the observed value y and the fitted value $\hat{y}$. Referring to Conte et al. (1986), MPAD is also known as the average measure for the magnitude relative error (MRE). The computation of MPAD for a particular curve is conducted at the location of w non-missing points as follows:

$$\text{MPAD} = \frac{1}{w}\sum_{j=1}^{w}\text{MRE} = \frac{1}{w}\sum_{j=1}^{w}\left|\frac{y_j - \hat{y}_j}{y_j}\right|. \tag{6}$$

The following is the procedure to conduct the FMeanBIM or FMedBIM imputation method. For an interval of time (1, 24) an incomplete data set of matrix **X** with the dimension of 153 days × 24 h is imputed row by row (i.e. by daily basis in the context of the data set employed in this study) according to the following procedure:

Replace the missing values at time (t) of the location point j with the corresponding point mean/ median value of the complete data set; Fit day data points to form a curve using the model given by (1) with the fitness criteria given in (2); Compute the MPAD for the fitted curve given the name of MPAD_1; Replace the value at the missing location at time t of point j with the value obtained on the fitted curve; Fit the day data again to form a curve using the model given by (1) with the fitness criteria given in (2). Compute the MPAD for the newly fitted curve given the name of MPAD_2; and If $|\ \text{MPAD}_2 - \text{MPAD}_1| \leq 0.005$, stop the imputation; else, repeat steps (5-7).

The missing values can be replaced by the imputed values that can be obtained at the same location time (t) of point j on the fitted curve resulted from the final iteration.

Roughness penalty without based value method (RPoBM), roughness penalty functional mean-based iterative method (RPFMeanBIM) and roughness penalty functional median-based iterative method (RPFMedBIM) Three imputation method designs that apply the model for roughness penalties are the iterative methods without initial value (RPoBIM), the iterative method with the functional mean-based/initial value (RPFMeanBIM) and the functional median-based/initial value (RPFMedBIM). The procedure to impute the missing value using the curve obtained by applying the roughness penalty framework is equivalent to the previous procedure when using the FMeanBIM or FMedBIM method. However, this time, the specified λ value is required. The following is the procedure to impute the missing value using RPoBM method.

Fit the incomplete day data using the model (1) with the fitness criteria given in (3) to estimate a function *x(t)*; Replace the missing value with the value obtained from the predicted curve at time t is missing; Fit the complete day data again using the model (1) with the fitness criteria given in (3); Compute the MPAD given the name of MPAD_1; Replace the value at the missing locations with the value obtained from the new predicted curve; Fit the day data again to form a curve using the model given by (1) with the fitness criteria given in (3); and Compute the MPAD for the newly fitted curve given the name of MPAD_2; If $|\ \text{MPAD}_2 - \text{MPAD}_1| \leq 0.005$, stop the imputation; else, repeat steps (5-8).

For the RPFMeanBIM and RPFMedBIM methods, the procedure was similar to the procedure for RPoBM. The only difference is that to estimate the function *x(t)* at step (1), the initial mean or median value must be allocated at the missing points before the data conversion takes place. The second, third and the rest of the steps follow the same approach.

PERFORMANCE EVALUATION OF THE IMPUTATION METHODS

In order to evaluate the performance of the imputation methods, three performance indicators, namely, the coefficient of determination (Rsquare), index of agreement (AI) and the mean absolute error (MAE) adopted from Junninen et al. (2004) are considered. R square measures the proportion of variance captured that is explained by the model while the square root of Rsquare indicates the strength of the relationship between the predicted and the observed value. Due to the inefficiency of Rsquare in determining the size of the discrepancies between the observed and the estimated values (Willmott et al.

1985) MAE is used. MAE provides a sensitive measure of the residual, the average error of the model whereas AI measures the agreement in terms of similarity between the observed and the predicted value.

For all indicators, the mean is computed over 100 indicator matrices. The evaluation process starts by purposely eliminating some percentages of the observed data at random from a set of complete data. The aim was to reproduce the missing pattern. Each missing data will be replaced with the imputed value obtained using the considered imputation methods. The imputed value is then compared with the observed or actual data.

Suppose that there are P numbers of imputed values with the p^{th} value is $\hat{y}_p$ and the corresponding actual value is y_p. The average of the actual data is $\bar{y}$ and the average of the imputed data $\bar{\hat{y}}$, with σ_y and $\sigma_{\hat{y}}$ are their standard deviations; thus, the computation of the performance indicators are according to the following formulae:

$$\mathrm{AI} = 1 - \left[\frac{\sum_{p=1}^{P}\left(y_p - y_p\right)^2}{\sum_{p=1}^{P}\left(\left|\hat{y}_p - \bar{y}\right| + \left|y_p - \bar{y}\right|\right)^2}\right]. \tag{7}$$

$$\mathrm{MAE} = \frac{1}{P}\sum_{p=1}^{P}\left|\left(y_i - \hat{y}_p\right)\right|. \tag{8}$$

$$\mathrm{Rsquare} = \left[\frac{\sum_{p=1}^{P}\left(\hat{y}_p - \bar{\hat{y}}\right)\left(y_p - \bar{y}\right)}{\sigma_y \sigma_{\hat{y}}}\right]^2. \tag{9}$$

Rsquare and AI take on the values between 0 and 1, with values closer to 1 imply a better fit. On the other hand, MAE ranges from 0 to infinity and a better fit is obtained when MAE approaches 0. To help in comparing the performance of the proposed imputation methods concurrently with the Rsquare and AI values, the MAE values are standardized into new performance indices, namely MBPI (MAE-based performance index). MBPI is constructed such that the values must lie within the range between 0 and 1 inclusively and the value closer to 1 is the ideal value. Letting $\mathrm{MAE}_{\max}$ and $\mathrm{MAE}_{\min}$ be the maximum and the minimum value of MAE, respectively, MBPI is computed using the following formula:

$$\mathrm{MBPI} = \frac{\mathrm{MAE}_{\max} - \mathrm{MAE}}{\mathrm{MAE}_{\max} - \mathrm{MAE}_{\min}}. \tag{10}$$

Results and Discussion

All seven imputation methods from three categories of functional-based imputation methods (single, iterative without roughness penalty approach and iterative with roughness penalty approach) have been tested on six different patterns at 5, 10 and 15% missing rate up to two types of maximum missing gap per day with the maximum of three and seven consecutive missing hours. The results of the overall performance of the methods are reported in Table 2.

Irrespective of the missing pattern, the ranking of performance in Table 2 shows that those methods with roughness penalty including the RPoBIM, RPFMeanBIM and RPFMedBIM give better results. It was also found that the iterative methods (FMeanBIM and FMedBIM) outperform the non-iterative approach. However, the best approach among these cannot be concluded yet based on the above summarized statistics. Further investigation on the accuracy measurement to compare the performance will be discussed. In terms of efficiency, all the iterative methods require on the average between one to two iterations to converge.

Figure 2(a)-2(c) presents box plots of the performance evaluation for the best three methods (RPoBIM, RPFMeanBIM and RPFMedBIM) in terms of MAE, R square and AI. On each box plot, the central mark is the median; the edges of the box are the 25 and 75th percentiles and the whisker represents the most extreme points, whereas outliers are plotted beyond the whiskers. The comparison of the methods is conducted on the 100 simulated data sets that have 5, 10 and 15% missing values and also with a maximum length of three and seven consecutive missing gaps.

Based on Figure 2(a), the average error (MAE) that represents the size of the discrepancies between the observed and the estimated values for the case of the missing values with up to three consecutive missing gaps is smaller than that of the seven consecutive missing gaps. RPFMeanBIM and RPFMedBIM methods perform better than RPoBIM method both at different missing percentages and different maximum gap lengths because they produce a lower error (represented by a smaller median) and lower variance (represented by a smaller size of box plot). The performances of the three methods are also found to be unaffected by the missing percentage but affected by the number of missing gaps. Figure 2(b) shows the accuracy as quantified by the amount of variance/information explained by the method by means of Rsquare. A higher value of Rsquare indicates better imputation. RPFMeanBIM and RPFMedBIM methods outperform RPoBIM method regardless of the missing percentage. The accuracy in terms of similarity measured by the agreement index (AI) between the observed and imputed values is shown in Figure 2(c). All three methods experience a decrease similar to the larger gap of missing values but the similarity is independent of the missing percentage. RPoBIM method performs better in the median but with a larger variation of the AI measurement compared to RPFMeanBIM and RPFMedBIM methods when the maximum gap length increases from three to seven.

TABLE 2. Performance of seven imputation methods

Pattern	Indicator	Imputation Methods						
		Non-iterative		Iterative		Roughness penalty		
		FMeanDM	FMedDM	FMeanBIM	FMedBIM	RPoBM	RPFMeanBIM	RPFMedBIM
P05_G3	MBPI	0.489	0.467	0.522	0.417	0.6214	0.593	0.587
	Rsquare	0.052	0.021	0.131	0.079	0.476	0.512	0.513
	AI	0.329	0.398	0.510	0.516	0.822	0.780	0.798
	Average	0.290	0.295	0.388	0.337	0.640	0.628	0.633
	Ranking	7	6	4	5	1	3	2
P05_G7	MBPI	0.504	0.551	0.526	0.547	0.605	0.571	0.595
	Rsquare	0.057	0.019	0.082	0.0372	0.281	0.323	0.324
	AI	0.334	0.393	0.422	0.445	0.701	0.629	0.625
	Average	0.298	0.321	0.344	0.343	0.529	0.508	0.515
	Ranking	7	6	4	5	1	3	2
P10_G3	MBPI	0.479	0.585	0.452	0.571	0.527	0.422	0.467
	Rsquare	0.050	0.017	0.122	0.068	0.492	0.520	0.521
	AI	0.332	0.405	0.505	0.514	0.832	0.803	0.801
	Average	0.287	0.336	0.360	0.384	0.617	0.582	0.596
	Ranking	7	6	5	4	1	3	2
P10_G7	MBPI	0.631	0.656	0.626	0.6820	0.5415	0.627	0.626
	Rsquare	0.061	0.019	0.086	0.0406	0.275	0.329	0.330
	AI	0.339	0.407	0.430	0.461	0.701	0.640	0.637
	Average	0.344	0.361	0.381	0.395	0.506	0.532	0.531
	Ranking	7	6	5	4	3	1	2
P15_G3	MBPI	0.468	0.477	0.459	0.478	0.630	0.608	0.637
	Rsquare	0.044	0.015	0.115	0.0631	0.461	0.498	0.497
	AI	0.343	0.4111	0.507	0.515	0.817	0.795	0.793
	Average	0.285	0.301	0.360	0.352	0.636	0.634	0.642
	Ranking	7	6	4	5	2	3	1
P15_G7	MBPI	0.607	0.573	0.595	0.581	0.612	0.476	0.495
	Rsquare	0.0489	0.016	0.075	0.038	0.253	0.320	0.322
	AI	0.328	0.401	0.419	0.456	0.683	0.637	0.632
	Average	0.328	0.333	0.363	0.358	0.516	0.478	0.483
	Ranking	7	6	4	5	1	3	2

COMPARING THE PERFORMANCE BETWEEN THE SINGLE AND ITERATIVE APPROACH

Based on Table 3, a comparison of the functional mean-based iteration method (FMeanBIM) to the single imputation method using functional average (FMeanDM) indicates improvements in the range from 1 to 6% for MAE and from 26 to 56% for AI. Meanwhile, the comparison between the median-based iteration method (FMedBIM) to the single imputation method using functional median (FMedDM) indicates improvements between 2 and 8% for MAE and between 11 and 30% for AI. The results of the study showed that the iterative method could improve the single imputation method. These results are supported by Zhang (2011).

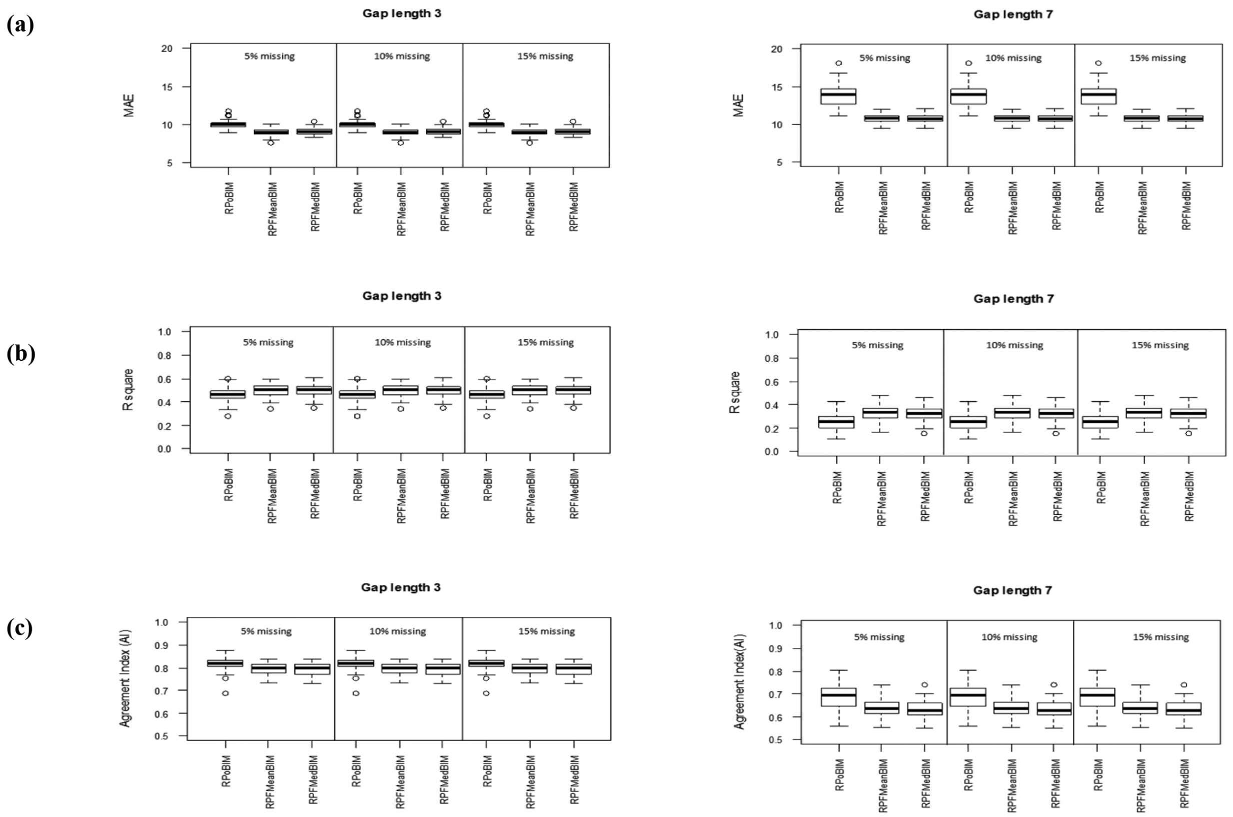

FIGURE 2. Accuracy measurement based on MAE (a), Rsquare (b) and AI (c) performance indicator

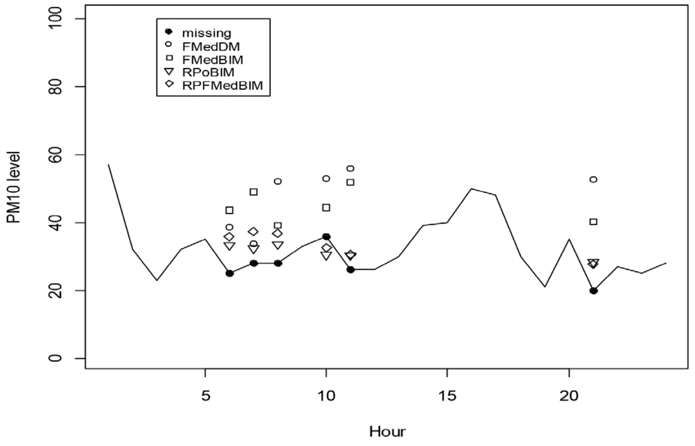

FIGURE 3. Example of the estimated missing points before and after missing treatment

COMPARING THE PERFORMANCE BETWEEN THE ITERATIVE WITHOUT AND WITH ROUGHNESS PENALTY APPROACH

The results in Table 4 shows that both RPFMeanBIM and RPFMedBIM methods indicates improvements in the range between 15 and 37% for MAE and between 27 and 59% for AI. Thus, the results provided the evidence that the roughness penalty was superior to the approach without roughness penalty.

For illustration purpose, the graphical representations of the imputation results are depicted in Figure 3. Consider the 24 hourly recorded data for one day. The solid circle point represents the data that is purposely eliminated. FMedDM (from the non-iterative method), FMedBIM (from the iterative method), RPoBIM (the iterative roughness penalty without initial based) and RPFMedBIM (the iterative roughness penalty with median initial value) are employed to estimate the missing points. The imputed points from the different methods are represented by different shapes; circle for FMedDM, square for FMedBIM, point-down triangle for RPoBIM and diamond

TABLE 3. Percentage of improvements of iterative imputation relative to single imputation methods

Method	MAE						AI					
	P05-G3	P05-G7	P10-G3	P10-G7	P15-G3	P15-G7	P05-G3	P05-G7	P10-G3	P10-G7	P15-G3	P15-G7
FMeanBIM	12.216	12.917	12.594	12.848	12.699	12.855	0.510	0.422	0.505	0.430	0.507	0.419
FMeanDM	12.863	13.190	13.213	13.074	13.311	13.081	0.329	0.334	0.332	0.339	0.343	0.328
Percentage improve	5.029	2.069	4.685	1.729	4.598	1.728	55.015	26.347	52.108	26.844	47.813	27.744
FMedBIM	13.599	14.667	14.197	14.394	14.375	14.293	0.516	0.445	0.513	0.462	0.515	0.456
FMedDM	14.662	15.145	15.156	14.793	15.348	14.682	0.398	0.393	0.405	0.407	0.411	0.401
Percentage improve differe	7.250	3.156	6.328	2.697	6.340	2.650	29.648	11.685	21.053	13.514	25.304	13.716

TABLE 4. Percentage improvements of roughness penalty-based iterative (with mean/median initial value) methods relative to non-roughness penalty

Method	MAE						AI					
	P05-G3	P05-G7	P10-G3	P10-G7	P15-G3	P15-G7	P05-G3	P05-G7	P10-G3	P10-G7	P15-G3	P15-G7
RPFMeanBIM	8.757	10.868	9.015	10.827	9.059	10.804	0.800	0.629	0.803	0.640	0.795	0.637
FMeanBIM	12.216	12.917	12.594	12.848	12.699	12.855	0.510	0.422	0.505	0.430	0.507	0.419
Percentage different	28.315	15.863	28.418	15.730	28.664	15.955	56.863	49.052	59.000	48.837	56.805	52.029
RPFMedBIM	8.744	10.812	9.012	10.796	9.061	10.762	0.516	0.445	0.513	0.462	0.515	0.456
FMedBIM	13.599	14.667	14.197	14.394	14.375	14.293	0.798	0.625	0.801	0.637	0.793	0.632
Percentage improve differe	35.701	26.283	36.522	24.996	36.966	24.704	35.338	28.800	35.955	27.473	35.057	27.848

for RPFMedBIM method. Figure 3 shows an interesting result; the imputed points from RPoBIM and RPFMedBIM methods are found closer to the real observed value.

CONCLUSION

In this study, seven imputation methods based on the application of FDA have been introduced to impute the missing values. The imputation methods are classified into three categories; single imputation, iterative imputation without roughness penalty and iterative imputation with roughness penalty approach. The results showed that the three best methods are from the roughness penalty approach with iterative imputation. The ranking of the performance from the best to the least are as follows: the first belongs to the iterative method with the roughness penalty approach; the second is the iterative method without roughness penalty approach; and the least is the single imputation approach.

All the methods are independent of the missing percentage but dependent on the size of the consecutive gap length. It is also found that the imputation using the iterative method produced an improvement in the performance compared with the non-iterative imputation method. On top of that, using the initial mean or median value in the imputation process provides an additional advantage for the real application. The values on the median or mean curve can be used as the estimated value at the missing points for days with too many missing hours or even for days with no recorded data. Overall, the iterative imputation method using the roughness penalty approach is identified to be more flexible and superior to other methods.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Environment Malaysia and Prof. Dr. Mohd Talib Latif from Universiti Kebangsaan Malaysia (UKM) for providing the information and data. This work is supported by the UKM's Research University Grant (UKM-AP-2011_19).

REFERENCES

Acuna, E. & Rodriguez, C. 2004. The treatment of missing values and its effect in the classifier accuracy. In *Classification, Clustering and Data Mining Applications,* edited by Banks, D., House, L., McMorris, F.R., Arabie, P. & Gaul, W. Berlin Heidelberg: Springer. pp. 639-648.

Baraldi, A.N. & Enders, C.K. 2010. An introduction to modern missing data analyses. *Journal of School Psychology* 48: 5-37.

Cao, Y., Poh, K.L. & Cui, W.J. 2008. A non-parametric regression approach for missing value imputation in microarray. In *Intelligent Information Systems XVI. Proceedings of the International IIS'08 Conference*. pp. 25-34.

Chen, J., Li, E., Lau, A., Cao, J. & Wang, K. 2010. Automated load curve data cleansing in power systems. *IEEE Transaction Smart Grid* 1(2): 213-221.

Conte, S.D., Dunsmore, H.E. & Shen, V.Y. 1986. *Software Engineering Metrics and Models*. Menlo Park, California, USA: The Benjamin/Cummings Publishing Company.

Craven, P. & Wahba, G. 1979. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numeriche Mathematik* 31: 377-403.

Gao, H.O. & Niemeier, D.A. 2008. Using functional data analysis of diurnal ozone and NOx cycles to inform transportation emissions control. *Transportation. Research Part D* 13: 221 - 238.

Huang, J.Z. & Shen, H. 2004. Functional coefficient regression models for non-linear time series: a polynomial spline approach. *Scandinavian Journal of Statistics* 31: 515-534.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. & Kolehmainen, M. 2004. Method for imputation of missing values in air quality data sets. *Atmospheric Environment* 38: 2895-2907.

Malek, M.A., Harun, S., Shamsuddin, S.M. & Mohamad, I. 2008. Reconstruction of missing daily rainfall data using unsupervised artificial neural network. *World Academic of Science Engineering and Technology* 44: 616-621.

Martinez, J., Saadvera, A., Garcia-Nieto, P.J., Pineiro, J.I., Iglesias, C., Taboada, J., Sancho, J. & Pastor, J. 2014. Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). *Applied Mathematics and Computation* 241: 1-10.

Park, A., Guillas, S. & Petropavlovsikh, I. 2013. Trends in stratospheric ozone profiles using functional mixed model. *Atmospheric Chemistry and Physics* 13: 11473-11501.

Pighin, M. & Ieronutti, L. 2008. A methodology supporting the design and evaluating the final quality of data warehouse. *Int. J. Data Warehouse Min*. 4(3): 15-34.

Plaia, A. & Bondi, A.L. 2006. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment* 40: 7316-7330.

Police, A. & Lasinio, G.J. 2009. Two approaches to imputation method of missing values in environmental pollution data sets. *Journal of Data Science* 7: 43-59.

Preda, C., Duhamel, A., Picavet, M. & Kechadi, T.I. 2005. Tools for statistical analysis with missing data: Application to a large medical database. In *Connecting Medical Informatics and Bio-Informatic Proceedings of MIE 2005*, edited by Engelbrecht, R., Geissbuhler, A., Lovis, C. & Mihalax, G. *ENMI*. pp. 181-186.

Quintela-del-Rio, A. & Francisco-Fernandez, M. 2011. Nonparametric functional data estimation applied to ozone data: Prediction and extreme value analysis. *Chemosphere* 82: 800-808.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsay, J.O. & Silverman, B.W. 2006. *Functional Data Analysis*. 2nd ed. New York: Springer.

Ramsay, J.O., Hooker, G. & Graves, S. 2009. *Functional Data Analysis with R and Mathlab*. New York: Springer.

Ruggieri, M., Plaia, A., Salvo, F.D. & Agro, G. 2013. Functional principal component analysis for the explorative analysis of multisite-multivariate air pollution time series with long gaps. *Journal of Applied Statistics* 40(4): 795-807.

Shaadan, N., Jemain, A.A., Latif, M.T. & Deni, S.M. 2015. Anomaly detection and assessment of PM10 functional data at several locations in the Klang Valley, Malaysia. *Atmospheric Pollution Research* 6: 365-375.

Shaadan, N., Deni, S.M. & Jemain, A.A. 2012. Assessing and comparing PM10 pollutant behavior using functional data approach. *Sains Malaysiana* 41(11): 1335-1344.

Smolinski, H. & Hlawiczka, S. 2007. Chemometric treatment of missing elements in air quality data sets. *Pollution Journal of Environmental Studies* 16: 613-622.

Torres, J.M., Nieto, P.J.G., Alejano, L. & Reyes, A.N. 2011. Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of Hazardous Material* 186: 144-149.

Wilmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J. & Rowe, C.M. 1985. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research* 90(C5): 8995-9005.

Zhang, S. 2011. Shell-neighbor method and its application in missing data imputation. *Applied Intelligent* 35: 123-133.

Norshahida Shaadan* & Sayang Mohd Deni
Center for Statistical and Decision Science Studies
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA (UiTM)
40450 Shah Alam, Selangor Darul Ehsan
Malaysia

Abdul Aziz Jemain
DELTA, School of Mathematical Sciences
Faculty of Science & Technology
Universiti Kebangsaan Malaysia (UKM)
43600 Bangi, Selangor Darul Ehsan
Malaysia

*Corresponding author; email: shahida@tmsk.uitm.edu.my

Received: 26 March 2014
Accepted: 15 June 2015